A Comparison of Individual-Based and Mean Field Model Approaches as Applied to Vector-Borne Pathogen Spread

Morganne Igoe

Submitted under the supervision of Allison Shaw to the University Honors Program at the University of Minnesota-Twin Cities in partial fulfillment of the requirements for the degree of Bachelor of Science, *summa cum laude* in Mathematics.

5/8/2017

I would like to thank my thesis advisor, Professor Allison Shaw, for all of her guidance and work on this project, as well as Professors Borer and Nykamp for their feedback on my work. I would also like to thank Evelyn Strombom, Lindsay Leverett, Dr. Lauren Sullivan, Kate Meyer, and Julie Sherman for their feedback throughout this process, as well as the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the research results reported within this paper.

Abstract

The choice of a modeling approach is a critical decision in the modeling process, as it determines the complexity of the model and the phenomena that the model captures. In this paper, we aim to determine how the construction and results of an individual-based model differ from those of the mean field model on which it is based. An individual-based model examining the effect of vector life history and behavior traits on vector-borne pathogen spread was constructed, based on the corresponding mean field model that was formulated and analyzed previously by others. The structure, results, and analysis of the two models were then compared. We found that the individual-based model produced slower population dynamics than the mean field model, due to the addition of stochastic components and clumping in the individual-based model. However, despite this difference in the speed of the dynamics, we see roughly the same results between the mean field model and the individual-based model. This shows us that, for this particular system, the additional complexity and stochasticity of the individual-based model did not yield any major insights into the overall behavior of the system that was not already captured by the mean field model, but does allow us to investigate the effect of smaller-scale dynamics, such as clumping, on the system.

Introduction

One of the most important decisions that needs to be made when modeling a system is choosing a modeling approach. The choice of modeling approach determines what level of detail is included in the model, and therefore determines which phenomena are captured by it. In Durrett and Levin [1], four different modeling approaches were applied to three biological systems, each representing a different set of biological assumptions, and their results were compared. In one of the systems, the long-term results differed greatly between spatial and non-

spatial models, where the spatial models showed the possibility of coexistence between two competing species, and the non-spatial models showed that one species would become extinct. In another biological system, the long-term results differed between stochastic models dealing with discrete individuals and differential equation models, where in the discrete, stochastic system, the coexistence of two competing species was possible, while in the differential equation model, both species became extinct [1]. This shows that the modeling approach used to analyze a system has a significant effect on the results, and it is therefore important to understand how the choice of modeling approach will impact the results. This paper will focus on formulating an individual-based model of the spread of barley yellow dwarf virus by aphids among plants and comparing the differences and similarities between it and the corresponding mean field model.

Mean field models are described using ordinary differential equations (ODE's). A mean field approach assumes that all individuals have the same probability of interacting with any other individual in the model, allowing us to simplify and study complex biological systems. However, there are limits to the phenomena that can be explored using a mean field model [1]. For example, because mean field models operate using ODE's, it is assumed that the density of individuals is the same across the entire model. This assumption does not allow for phenomena such as clumping, where the density of individuals in one spatial area of the model increases suddenly due to reproduction in that area [2]. In a model where events such as reproduction or movement are density-dependent, this local increase in density can be important to local population dynamics, but cannot be modeled using a mean field model, making another modeling technique possibly more appropriate.

Individual-based modeling is a modeling technique that simulates the behavior of individuals in a system to study both how the system affects the individuals and how the

individuals affect the system [3]. Individual-based models (IBMs) allow for variation in the characteristics of the individuals in the model, such as fecundity and mortality. These characteristics are adaptive, meaning they change based on the state of the individual, as well as the state of the surrounding individuals and the environment. This individuality and adaptiveness allows us to investigate which system dynamics emerge due to the interactions between individuals and their environment [4].

The Biological System

Barley yellow dwarf virus (BYDV) is widely considered to be the most widespread and economically important disease affecting cereals [5]. It affects over 100 species of the Poaceae family, including crops such as wheat, barley, maize, and oats, all of which are important food and economic resources across the globe. BYDV is spread by aphids, which transmit the virus to and receive the virus from their host plants in the process of feeding on their hosts. A non-viruliferous aphid (an aphid not carrying the virus) can become viruliferous (able to transmit the virus to non-infected plants) by feeding on an infected plant, while a non-infected plant can become infected if it is fed on by a viruliferous aphid [5]. Since this virus is spread solely by vectors (in this case, aphids), it is important to understand which vector behavior traits influence pathogen spread the most, so that we can devise more effective methods for controlling the spread of this virus among crops.

The purpose of the individual-based model constructed here is to understand the effect of vector behavior and life history traits on the spread of barley yellow dwarf virus among wheat plants. The model is adapted from the differential equation model of persistently transmitted BYDV on wheat developed by Shaw et al. (See Appendix for model equations, hereafter referred to as the 'ODE model') [6]. An individual-based model allows for the inclusion of biologically

realistic characteristics of population growth that are not possible in a mean field model, including phenomena like clumping. This type of phenomenon is naturally included in the individual-based model described in this paper, allowing us to investigate its importance to local and global dynamics. We will compare the formulation of the two models and their results to determine how the similarities and differences in the models affect the outcome.

Model

The individual-based model is described below using the Overview, Design concepts, and Details (ODD) protocol outlined in Railsback and Grimm [3].

State variables

The model consists of two classes of individuals: individual vectors, who carry the pathogen, and individual hosts, who become infected with the pathogen and can transmit it to non-viruliferous vectors. Each individual vector has the following characteristics: ID number, infection status (non-viruliferous or viruliferous), location (given by the ID number of the host on which the vector is located), probability of departing from its host at the current time step, probability of becoming infected by its host at the current time step, and probability of reproducing at the current time step. Each individual host has the following characteristics: ID number, infection status (healthy or infected), number of vectors on the host (T), number of viruliferous vectors on the host (V), number of non-viruliferous vectors on the host (N), and probability of becoming infected by a viruliferous vector.

Initialization

The vector population is divided into four subclasses: non-viruliferous vectors on healthy hosts (N_h), viruliferous vectors on healthy hosts (V_h), non-viruliferous vectors on infected hosts

(N_i), and viruliferous vectors on infected hosts (V_i). Initially, the vector population is set to 10,000 individuals, 95% of which are of the subclass N_h and 5% of which are V_h. Vectors are assigned an ID number, as well as a random location at the beginning of the simulation. They are also assigned a reproduction, departure, and settlement probability according to the formulas described in the processes and scheduling section below. They are also assigned a probability of becoming infected.

The host population is divided into two subclasses: infected and healthy. The host population is set at 125,000 (the field size F), all of which are initially healthy. The host population does not undergo reproduction or maturation, and will remain the same size throughout the simulation. They are assigned an ID number, as well as a count of the total number of non-viruliferous and viruliferous vectors on the host. They are also assigned a probability of becoming infected by a viruliferous vector according to the formula described in the Processes and Scheduling section below.

The fraction of hosts that are infected (I) and healthy (H), as well as the number of non-viruliferous vectors on healthy hosts (N_h), the number of non-viruliferous vectors on infected hosts (N_i), the number of viruliferous vectors on healthy hosts (V_h), and the number of viruliferous vectors on infected hosts (V_i) are recorded at each time step, starting with time t=0.

Model Variables and Parameters

Variables	Definition	Units
N_h	Non-viruliferous vectors on healthy hosts in population	# vectors
N_i	Non-viruliferous vectors on infected hosts in population	# vectors
V_h	Viruliferous vectors on healthy hosts in population	# vectors

V_i	Viruliferous vectors on infected hosts in population	# vectors
N	Non-viruliferous vectors on an individual host	# vectors
V	Viruliferous vectors on an individual host	# vectors
T	Total number of vectors on an individual host	# vectors
H	Fraction of hosts that are healthy	-
I	Fraction of hosts that are infected	-

Parameters	**Definition**	**Units**	**Value**	**Range**
r_h	Intrinsic vector growth rate on healthy hosts	1/day	0.186	0.01-0.6
r_i	Intrinsic vector growth rate on infected hosts	1/day	0.263	0.01-0.6
K_h	Carrying capacity of vectors on healthy hosts	vectors/host	500	100-1000
K_i	Carrying capacity of vectors on infected hosts	vectors/host	500	100-1000
F	Field density	hosts/hectare	125,000	75000-125000
ε	Preference of viruliferous vectors for settling on infected hosts	hosts	0.7649	0.25-4
δ	Preference of non-viruliferous vectors for settling on healthy hosts	hosts	0.9376	0.25-4
a_h	Max departure rate of vectors form healthy hosts	1/day	0.1268	0.014-1.4
a_i	Max departure rate of vectors from infected hosts	1/day	0.1438	0.014-1.4
c_1	Same status half-departure constant	vectors/hosts	13.76	1.37-137
c_2	Different status half-departure constant	vectors/hosts	$0.3c_1$	1.37-137
β_v	Rate vector on infected host becomes viruliferous	1/day	1	0.2-1
β_i	Rate healthy hosts become infected	1/day	0.182	0.7-1

Table 1: Model Variables and Parameters. Parameter values and ranges are taken from [6].

Processes and Scheduling

The model runs in time steps of days. During each time step, each vector undergoes reproduction, feeding, departure and settlement. Each process is completed for all vectors before the next process is carried out.

Reproduction

We assume that the growth of the vector population is density dependent, therefore the probability of reproduction for each vector depends on the total number of vectors on its host (T). We also assume that the maximum growth rate of a vector is determined by the infection status of its host and that offspring produced in each time step inherit the infection status of their host, as well as the location of its parent. Vectors on healthy hosts reproduce with a maximum per capita growth rate of r_h, while vectors on infected hosts reproduce with a maximum rate of r_i. We assume that r_i is greater than r_h because the infection of a host with BYDV alters the host's quality, making nutrients more readily available to vectors on an infected host [7]. The carrying capacity of each host is assumed to be dependent on its infection status, where K_h is the carrying capacity of a healthy host and K_i is the carrying capacity of an infected host.

Therefore, for vectors on a healthy host, reproduction probability is given by the logistic function $r_h\left(1-\frac{T}{K_h}\right)$. For vectors on an infected host, reproduction probability is determined by the function $r_i\left(1-\frac{T}{K_i}\right)$.

To determine if a vector will reproduce, the reproduction probability for each host is compared to a randomly generated number between 0 and 1. If the random number is less than the probability, the vector will produce one offspring. If the random number is greater than or equal to zero, the vector will not reproduce during this time step.

After the vectors reproduce, the total number of vectors on each host, as well as the number of viruliferous and non-viruliferous vectors on each host are updated. The departure, reproduction, and infection probabilities for each vector, as well as the infection probability for each host, are also updated.

Feeding and Transmission

When a vector feeds on its host, there are two possible transmission events that can occur: a non-viruliferous vector becoming viruliferous by feeding on an infected host and a healthy host becoming infected by being fed on by a viruliferous vector. The probability of a non-viruliferous vector on an infected host becoming viruliferous is determined solely by the transmission rate β_v. The infection probability for all other vectors is set to zero. To determine if a non-viruliferous vector on an infected host will become infected, the infection probability for each vector is compared to a randomly generated number between 0 and 1. If the random number is less than the probability, the vector becomes viruliferous and its infection status is updated. If the random number is greater than or equal to the probability, the vector remains non-viruliferous.

The probability of a healthy host becoming infected from a viruliferous vector is density dependent and is determined by the formula $\beta_i * V$, where β_i is the rate healthy hosts become infected by a viruliferous vector and V is the number of viruliferous vectors on that host. To determine if a healthy host will become infected, the infection probability for each host is compared to a randomly generated number between 0 and 1. If the random number is less than the probability, the host becomes infected and its infection status is updated. If the random number is greater than or equal to the probability, the host remains healthy.

After the vectors feed on the hosts and the infection statuses of the vectors and hosts are updated, the number of viruliferous and non-viruliferous vectors on each host are updated, as well as departure, settlement, reproduction and infection probabilities for each vector and host.

Departure

For a vector to move between hosts, the vector must first depart from its current host, and then settle on a new one. A vector's probability of departure is dependent both on its own infection status (non-viruliferous or viruliferous) and the infection status of its host (healthy or infected). The maximum departure rate from a healthy host is represented by a_h, while the maximum departure rate from an infected host is represented by a_i. a_i is assumed to be greater than a_h. This is due to findings which suggest that plants alter vector behavior through the release of chemicals to enhance pathogen spread, leading to a higher departure rate for vectors found on infected plants [7]. The probability of a non-viruliferous vector departing from a healthy host is determined by the formula $\frac{a_h T}{c_1+T}$. The probability of a non-viruliferous vector departing from an infected host is determined by $\frac{a_i T}{c_2+T}$. The probability of a viruliferous vector departing from a healthy host is determined by $\frac{a_h T}{c_2+T}$. The probability of a viruliferous vector departing from an infected host is determined by $\frac{a_i T}{c_1+T}$. The parameters c_1 and c_2 are half departure constants, which represent the density of vectors that cause the departure rate to become half of the maximum departure rate. c_1 corresponds to vectors who have the same infection status as their host, and c_2 corresponds to vectors who have the opposite infection status of their host.

To determine if a vector will depart from its host, the departure probability for each vector is compared to a randomly generated number between 0 and 1. If the random number is

less than the departure probability, the vector will depart from its host, and if the random number is greater than or equal to the departure probability, the vector will remain on its current host.

Settlement

The probability of a vector landing on a host depends on the infection status of the host and the vector, as well as the fraction of the host population which is healthy (H) and infected (I). It is assumed that vectors have a preference for the type of host on which they will settle. These preferences are expressed by the parameters δ and ε, where δ is the preference of a non-viruliferous vector and ε is the preference of a viruliferous vector. When δ<1, the non-viruliferous vectors prefer to land on an infected host, and when δ>1, the non-viruliferous vectors prefer to land on a healthy host. If δ=1, the non-viruliferous vectors have no host preference. When ε<1, the viruliferous vectors prefer to land on a healthy host, and when ε>1, the viruliferous vectors prefer to land on an infected host. If ε=1, the viruliferous vectors have no host preference.

The probability of a non-viruliferous vector landing on a healthy host (p_{Nh}) is determined by $1 - I^{\delta}$. The probability of a non-viruliferous vector landing on an infected host (p_{Ni}) is determined by I^{δ}. The probability of a viruliferous vector landing on a healthy host (p_{Vh}) is determined by H^{ε}. The probability of a viruliferous vector landing on an infected host (p_{Vi}) is determined by $1 - H^{\varepsilon}$.

To determine whether a departing non-viruliferous vector will land on a healthy or infected host, the p_{Ni} value is compared to a randomly generated number between 0 and 1 for each non-viruliferous vector which has departed from its host. If the random number is less than

the p_{Ni} value, the vector will land on a randomly chosen infected host. If the random number is greater or equal to the p_{Ni} value, the vector will land on a randomly chosen healthy host.

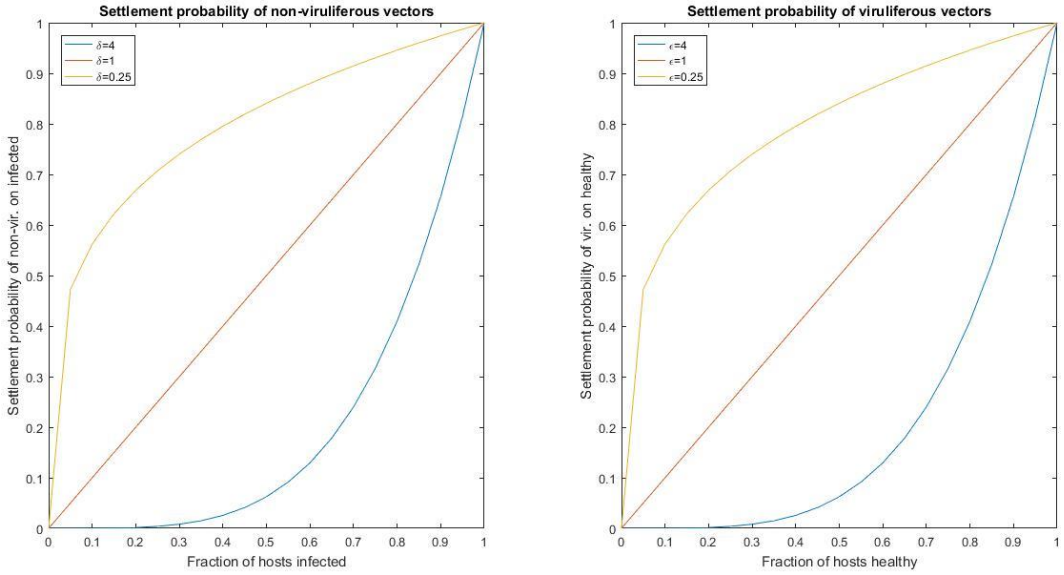

Figure 1: Left-hand plot: Plot showing the settlement probability of a non-viruliferous vector landing on an infected host, in terms of the fraction of hosts that are infected, for chosen delta values. Right-hand plot: Plot showing the settlement probability of a viruliferous vector landing on a healthy host, in terms of the fraction of hosts that are infected, for chosen epsilon values.

To determine whether a departing viruliferous vector will land on a healthy or infected host, the p_{Vh} value is compared to a randomly generated number between 0 and 1 for each viruliferous vector which has departed from its host. If the random number is less than the p_{Vh} value, the vector will land on a randomly chosen healthy host. If the random number is greater or equal to the p_{Vh} value, the vector will land on a randomly chosen infected host.

After the vectors disperse and settle, the total number of vectors on each host, as well as the number of viruliferous and non-viruliferous vectors on each host, are updated. The departure, infection and reproduction probabilities for each vector and host are also updated.

Differences from the ODE Model

Limitations due to the nature of IBM's

The model constructed by Shaw et al. [6] is a mean field model, which is described using ordinary differential equations. Therefore, the model is continuous in time and in population size, meaning that the model allows for the existence of fractions of days and individuals. Due to the nature of individual-based models, the model constructed here does not allow for fractions of individuals. The IBM is also iterative and proceeds in discrete time steps of days, and therefore does not allow for fractions of days in its calculations.

Because we must track and store a large number of characteristics for each vector and host, the number of vectors and hosts capable of being simulated is limited by computing power and memory storage. Therefore, the field density parameter F, which describes the number of hosts per hectare of land, was lowered to a fraction (1/32) of the F value used in the ODE model.

Parameter Calculations

The ODE model is a deterministic model, meaning that, given the same initial conditions and parameter values, the model will produce the same results each time. The individual-based model, on the other hand, contains stochastic components, meaning that the same set of initial conditions and parameter values may produce slightly different results each time the simulation is run. This stochasticity emerges due to how we adapted the ODE model, which is described in more detail below.

In order to adapt the ODE model into an individual-based model, values such as departure, reproduction and transmission rates were translated into probabilities. This is due to the fact that IBM's cannot simulate fractions of individuals. Therefore, the rate at which

individuals undergo a certain event, such as departure from their hosts, was translated into a probability of each individual making the decision to undergo that task, in order to remove the possibility of fractions of individuals appearing in the model. This adaptation allows the IBM to utilize the differential equations from the ODE model for calculating the probabilities of each event occurring, and also allows for individual-level variation in these traits, which is not possible using a mean field model.

In the ODE model, the departure rates of the vectors, the rate at which hosts became infected with the pathogen, and the reproduction rates are dependent on the density of vectors on each host. Since the ODE model is a mean field model, the density of vectors on each host is an average over the population of vectors. However, because the individual-based model keeps track of how many vectors are present on each individual host at each time step, the departure, reproduction and transmission probabilities in the IBM are calculated using the actual number of vectors on each host, as opposed to an average density of vectors on each host. This allows us to discard one of the parameters from the ODE model, namely the parameter ξ, the minimum hosts needed to support the vectors (See Appendix).

For example, the reproduction rate of vectors in the ODE model is assumed to be density dependent, where the per capita growth rate for vectors on infected hosts is determined by the logistic growth equation $r_i * \left(1 - \frac{N_i+V_i}{K_i*I*F}\right)$, where r_i is the instantaneous growth rate of vectors on an infected host, N_i is the number of non-viruliferous vectors in the population, V_i is the number of viruliferous vectors in the population, K_i is the carrying capacity of an infected host, I is the fraction of hosts that are infected, and F is the number of hosts per hectare of land. This growth rate is calculated based on the average density of vectors on infected hosts. For the individual based model, we are able to adapt this equation because we know the exact number of vectors on

each individual host. This makes the equation used to determine reproduction probability for vectors on infected hosts in the IBM $r_i * \left(1 - \frac{T}{K_i}\right)$, where T is the total number of vectors on the infected host where the vector is located. This number is then used as the probability of the vector of reproducing a single offspring. This is equivalent to the process used in the ODE model, where every individual produces that number of offspring, but eliminates the possibility of producing a fraction of an individual. This process was repeated for transmission, departure and settlement rates, resulting in the equations described in the Process and Scheduling section above. The ODE model equations from which these formulae were adapted are shown in the Appendix.

Simulations

For the following simulations, the initial host population is entirely healthy (H=1, I=0), and the vector population is set at 10,000, 5% of which are viruliferous. The field density F, which represents the number of hosts per hectare, is set to 125,000. All other parameter values come from Table 1. The simulations described below were run for both the IBM and the ODE models using these initial conditions so that the results of the two models could be compared.

Population Dynamics

First, we studied the dynamics of the host and vector populations, using the initial conditions described above. The fraction of hosts that are infected and healthy, as well as the populations of the four vector subclasses (N_h, N_i, V_h, V_i), were studied over a 30-day simulation (Figure 2, Figure 3). A heat map showing the clumping of vectors on each host was also created (Figure 4).

Sensitivity Analysis

The sensitivity analysis for the IBM was done using the same method and code as in the Shaw et al. paper [6]. The analysis was performed using Latin Hypercube Sampling (LHS) with the statistical Partial Rank Correlation Coefficient (PRCC) technique, which is effective when the model parameters have a nonlinear and monotonic relationship to the output [8]. LHS is a technique which involves sampling without replacement a set of model parameter combinations from preset ranges on these parameter values. The model is then run for each of these parameter combinations, letting us get an estimate of the average output with a limited number of samples. The output measure for the IBM is the time at which 80% of the hosts become infected.

The sensitivity of the output measure to a parameter is determined by the magnitude of the parameter's PRCC value, as well as its corresponding p-value, where p-value<0.05 is considered significant. A parameter with a PRCC value closer to +1 or -1 more strongly influences the output measure, where a negative PRCC value represents an inverse relationship to the output measure and a positive PRCC represents a direct relationship to the output measure. Our output measure for both the IBM and ODE models is the time until 80% of hosts become infected. This means that a negative PRCC value corresponds to a shorter time until 80% of hosts become infected, meaning the rate of infection has increased, and vice versa for a positive PRCC value. The sensitivity analysis was performed for the IBM and ODE models (Figure 5, Figure 8).

Because this model is individual-based, there are additional limitations to the LHS sensitivity analysis. Marino et al. 2008 discusses these limitations, which emerge due to the additional stochasticity that is introduced in individual-based models by way of probabilistic decision-making of the individuals. This additional stochasticity is not present in deterministic

models, such as the model described in Shaw et al. [6], and is difficult to take into account when performing LHS sensitivity analysis, because the technique was developed for deterministic models [8].

When building a model, there are two kinds of uncertainty to consider: aleatory uncertainty, which emerges due to additional stochastic components of the model, and epistemic uncertainty, which emerges due to uncertainty in model parameter values [8]. Epistemic uncertainty is often characterized as uncertainty which can be reduced through further data or further honing of the model, while aleatory uncertainty is characterized as uncertainty that cannot be reduced further [9]. In this model, examples of aleatory uncertainty include the processes that vectors undergo in order to reproduce, transmit the virus, and depart and settle from their hosts, all of which are probabilistic decisions. Due to the increase in aleatory uncertainty in individual-based models, it becomes difficult to untangle the effects of aleatory versus epistemic uncertainty on the output measure when performing sensitivity analysis [8].

Marino et al. 2008 discuss various ways to account for aleatory and epistemic uncertainty when performing sensitivity analysis on individual-based models. One method is a replication and averaging scheme, which happens in two steps. First, the model simulations are repeatedly run for each parameter combination produced by the LHS (replication). Then, the PRCC and p-values are calculated for each parameter combination using the average output measure of the set of simulations run for that parameter combination (averaging). This replication and average process is done in an attempt to limit the effect of aleatory uncertainty on the sensitivity analysis for the individual-based model. This process is discussed in further detail in Marino et al. 2008 [8]. We intend to perform sensitivity analysis using this replication and averaging scheme and to

compare the results to the sensitivity analysis without these extra correction measures to determine the effect of this additional aleatory uncertainty on the results (Figure 7).

Parameter Simulations

We also ran simulations where we varied one parameter type while keeping all other parameters constant, and analyzed the effects of these different parameter combinations on the time until 80% of hosts became infected. This allows us to analyze the importance of the various parameter types on the rate of infection (Figure 9-Figure 12).

Control Measure Analysis

Additionally, we ran simulations investigating possible control measures for the spread of the pathogen in an infected field. An individual-based model allows us to simulate situations where specific individual hosts and vectors are removed from the system, an analysis which is not possible when using a mean field model. We therefore ran simulations where we removed a certain number of infected hosts, as well as the vectors on those hosts, at each time step to determine how many infected hosts must be removed from a field to stop the spread of the infection (Figure 13, Figure 14). These simulations were run with the same initial conditions as described at the beginning of this section.

Results

Populations Dynamics

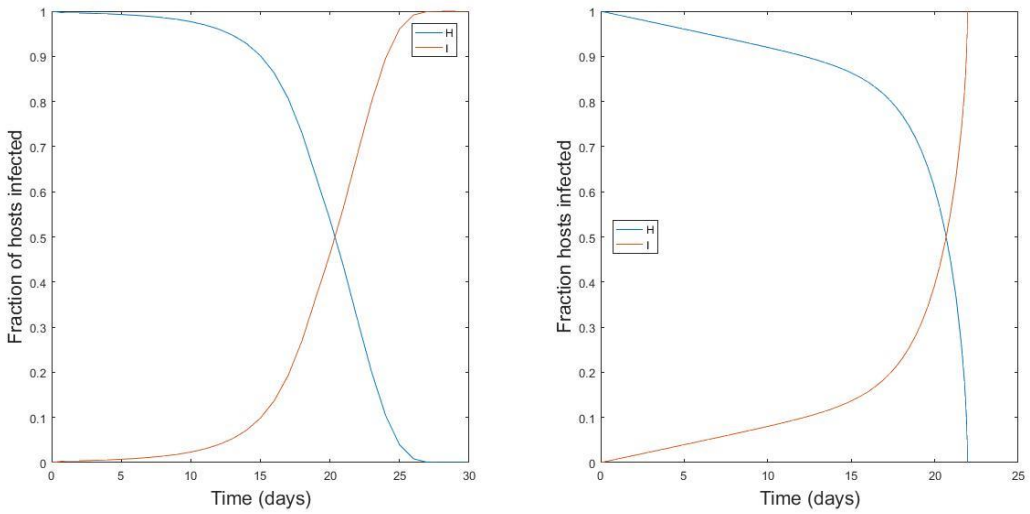

Figure 2: Left hand plot: Host dynamics over a 30 day simulation using the IBM model. Right hand plot: Host dynamics over a 30 day simulation using the ODE model. Initial Conditions: 5% infected, parameter values found in the parameter table in the Model Parameters section.

The dynamics of the host population are shown in Figure 2. Here we can see that the fraction of the host population that is infected (the red line) remains low until approximately 15 days have passed, at which point the fraction of hosts infected increases more rapidly. We also observe that it takes several days for the last of the hosts become infected.

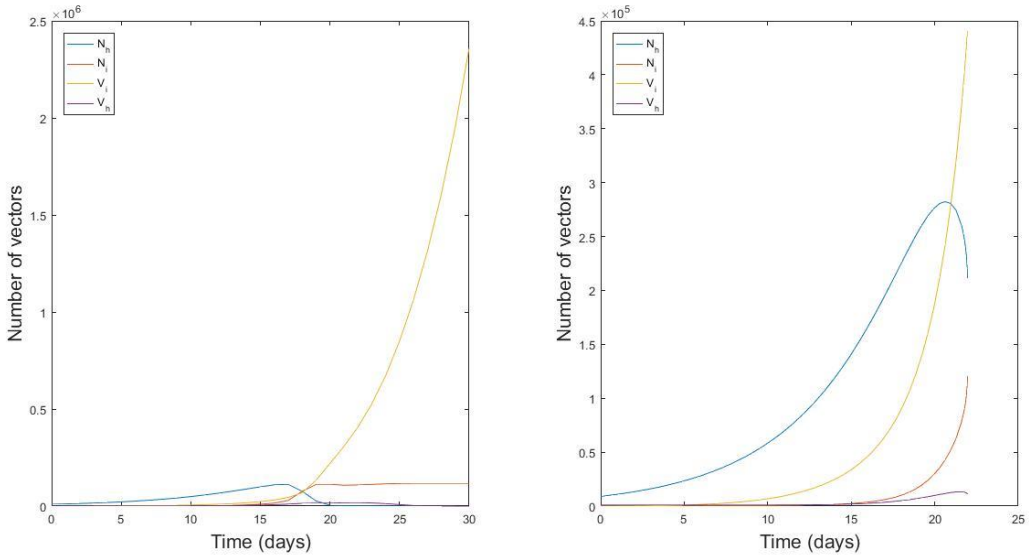

Figure 3: Left hand plot: Vector dynamics over a 30 day simulation. Right hand plot: Vector dynamics over a 30 day simulation using the ODE code. Initial Conditions: 5% infected, parameter values found in the parameter table in Table 1.

The dynamics of the vector population are shown in Figure 3, where the host population is divided into four categories: non-viruliferous vectors on healthy hosts (N_h), non-viruliferous vectors on infected hosts (N_i), viruliferous vectors on infected hosts (V_i), and viruliferous vectors on healthy hosts (V_h), all of which are expressed in number of vectors. We can see that initially the N_h population is the largest population class, but after just over 15 days, the size of the population class begins to decrease rapidly. This occurs as more hosts become infected, thereby increasing both the number of viruliferous vectors in the population and the number of non-viruliferous vectors on infected hosts in the population. This contributes to the sharp rise in the V_i population, as well as the more gradual increase in the V_h population and the leveling off of the N_i population. We would expect that the N_i population would begin to decrease over time after 30 days because the entire host population is infected by 30 days, leading to total infection of the N_i population over time. This is because no new non-viruliferous vectors could be produced after 30 days, as the vectors produced through reproduction inherit the infection status of their host, and all hosts are infected at 30 days.

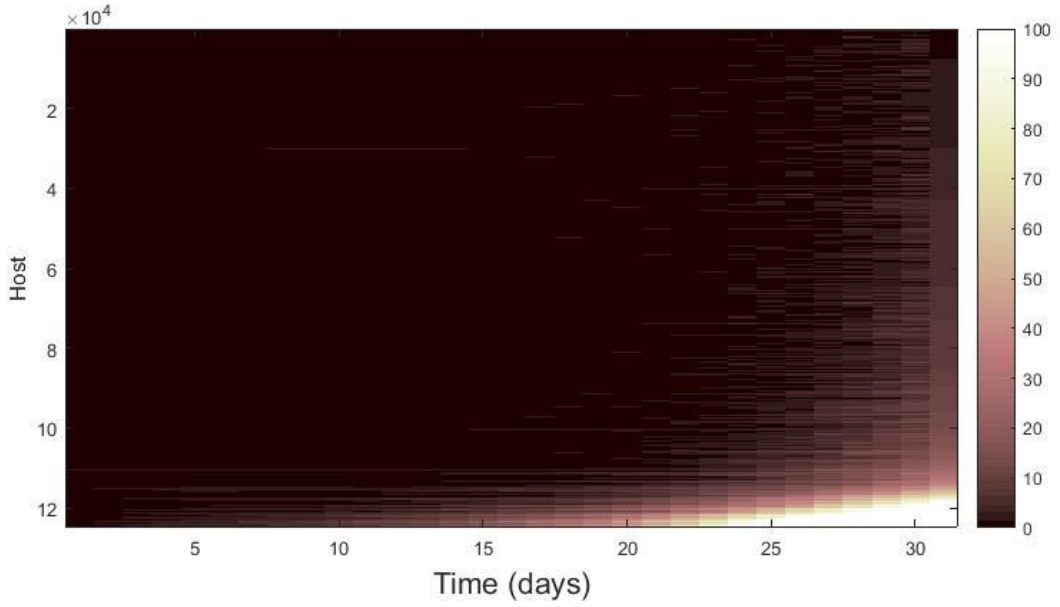

Figure 4: Heat map showing the clumping of vectors on each host. Each horizontal line represents a single host, and the color of the line at each time point represents the number of vectors on that host at that time point. A darker color represents a lower number of vectors on that host.

We see that a majority of hosts have very few vectors after 30 days, denoted by darker bars, and that only a small number have a large number of vectors, denoted by lighter bars (Figure 4). This shows the clumping of vectors on hosts that is possible with an individual-based model. The effect of this clumping on the model is further explored in the Sensitivity Analysis section below.

Sensitivity Analysis

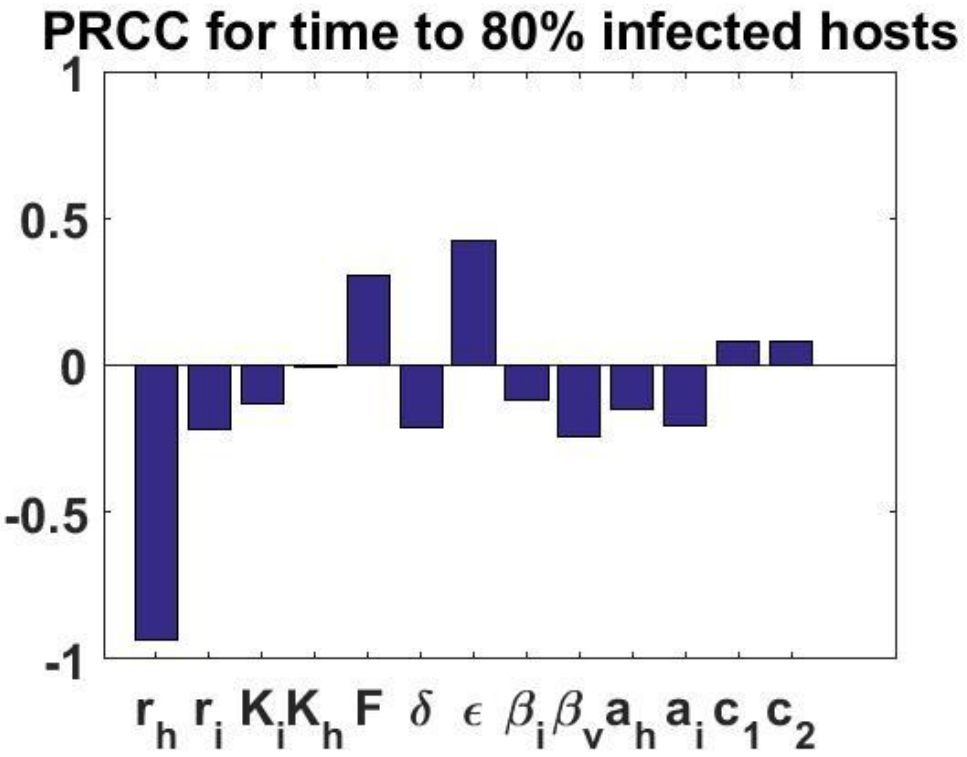

Figure 5: The results of the sensitivity analysis of the IBM after 500 samples using the LHS technique, which was also used in the Shaw et al. paper. Parameters which have a p-value>0.05 are considered non-significant, and include the following parameters: K_h, c_1, and c_2. All other parameter values were considered statistically significant, based on their p-values.

We see in the sensitivity analysis for the IBM in Figure 5 that the growth rate of vectors on healthy hosts (r_h) has the highest PRCC value, which is much larger than the PRCC value for r_i, the growth rate of vectors on infected hosts. We also see that β_v, the rate at which non-viruliferous vectors become viruliferous, has a higher PRCC value than β_i, the rate at which healthy plants become infected.

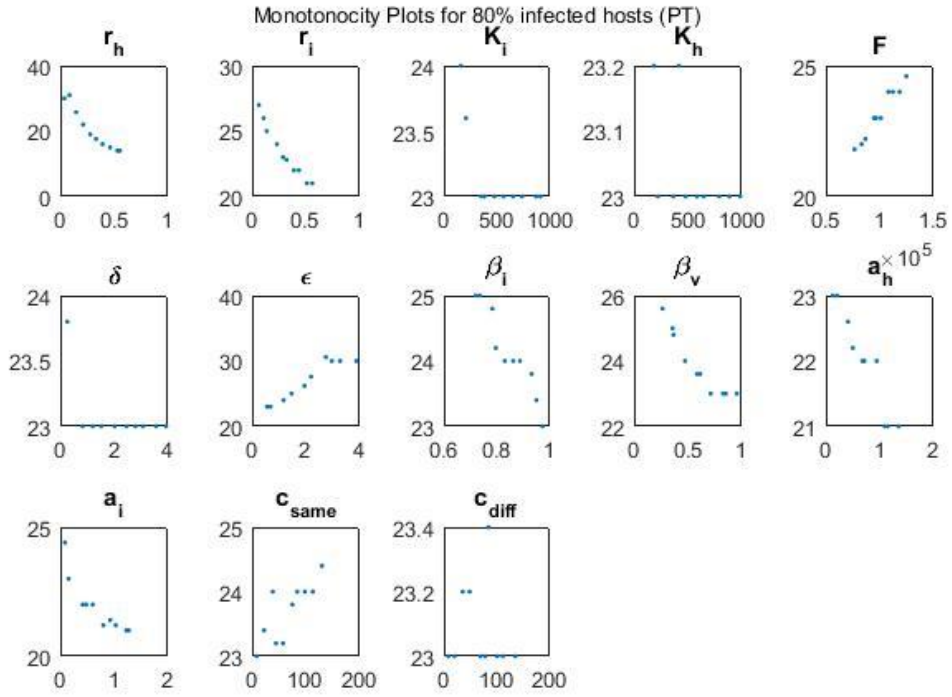

Figure 6: Plots showing the monotonic relationship between each parameter in the IBM model and the output measure, the time until 80% of hosts become infected. To account for the stochastic nature of the IBM, the simulations were repeated five time for each parameters value tested, and the results were averaged and reported above. In this plot, c_{same} refers to c_1 and c_{diff} refers to c_2.

Figure 6 shows the monotonic relationship between each parameter and the output measure. In the case of an IBM, we consider the relationship to be sufficiently monotonic to use LHS sensitivity analysis if the relationship between the parameter and the output measure is "fairly monotonic". In this case, we can see that all parameters have a fairly monotonic relationship with the output measure (Figure 6).

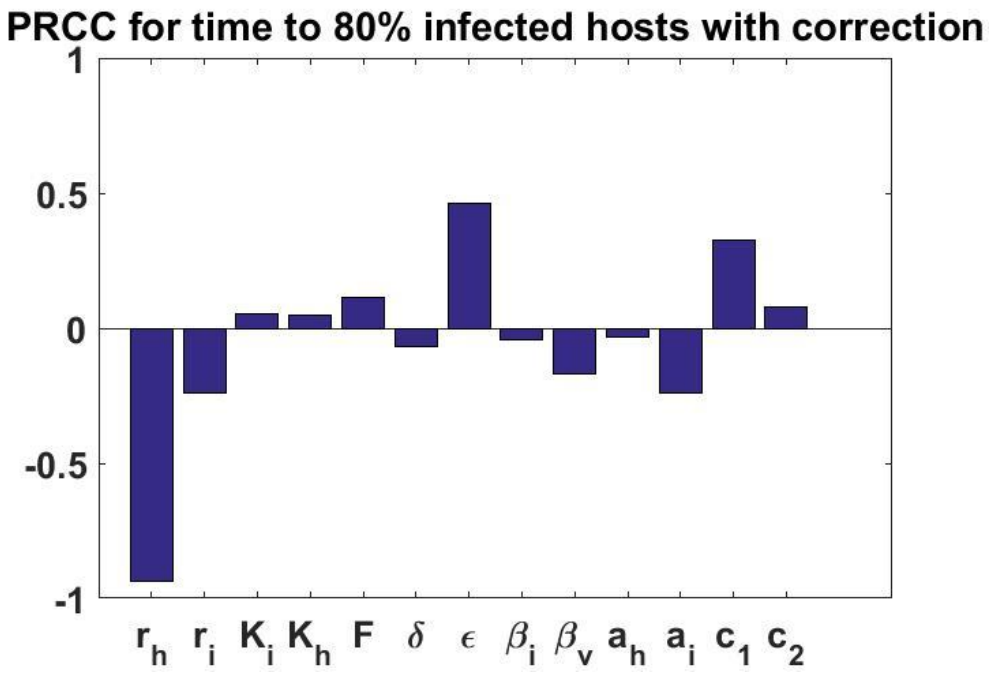

Figure 7: The results of the sensitivity analysis using the replication and averaging scheme described in Marino et al. 2008. The analysis was done for 100 samples. Each sample was simulated 5 times and then the results were averaged to produce the values used in the PRCC analysis. Parameters which have a p-value<0.05 are considered significant, and include the following parameters: r_h, r_i, ε, a_i, c_1. All other parameters were considered non-significant, according to their p-values.

When we perform sensitivity analysis using the replication and averaging scheme, we see the same relationship between the growth rates and infection rates, but we do see slightly different magnitudes for the PRCC values for some parameters (Figure 7). However, these differences may be attributed to the fact that we were only able to run the PRCC with replication for 100 samples, instead of 500, as we did in the original sensitivity analysis in Figure 5.

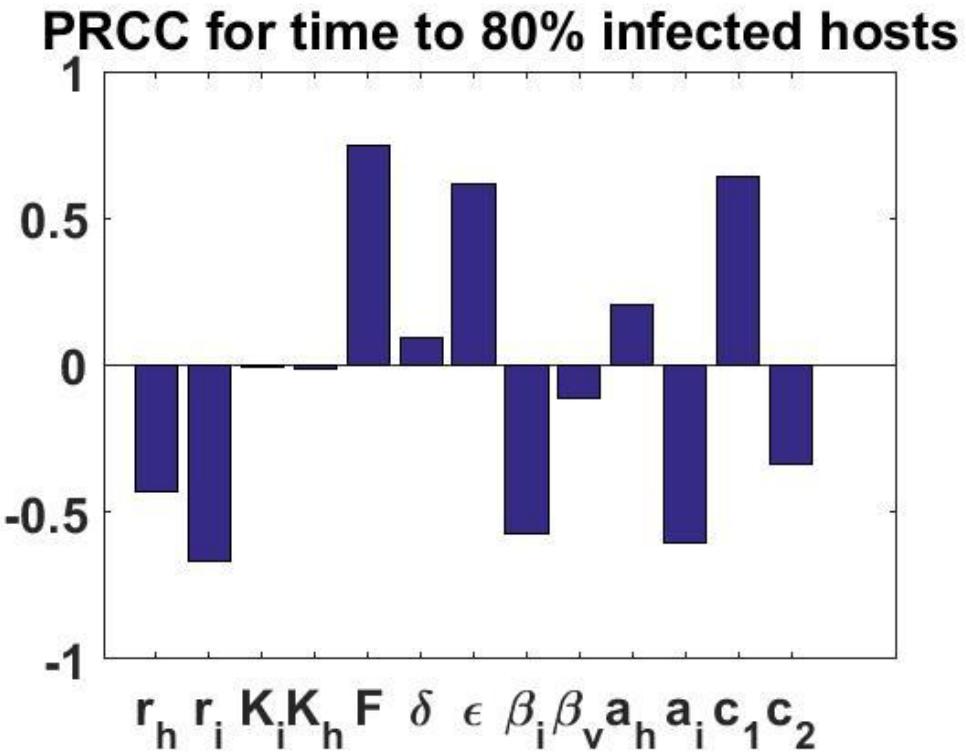

Figure 8: The results of the sensitivity analysis for the ODE model, run for 10,000 samples using the same parameters used in the IBM. This graph is provided for comparison to Figure 5 and Figure 7. Parameters which have a p-value>0.05 are considered non-significant, and include the following parameters: K_h and K_i. All other parameter values were considered statistically significant, based on their p-values.

In the sensitivity analysis for the ODE model in Figure 8, we see the opposite relationship between r_h and r_i and β_v and β_i than we did in the IBM. We also see that the r_h and r_i parameters have smaller PRCC values in the ODE model than they do in the analysis of the IBM model in Figure 5.

Parameter Simulations

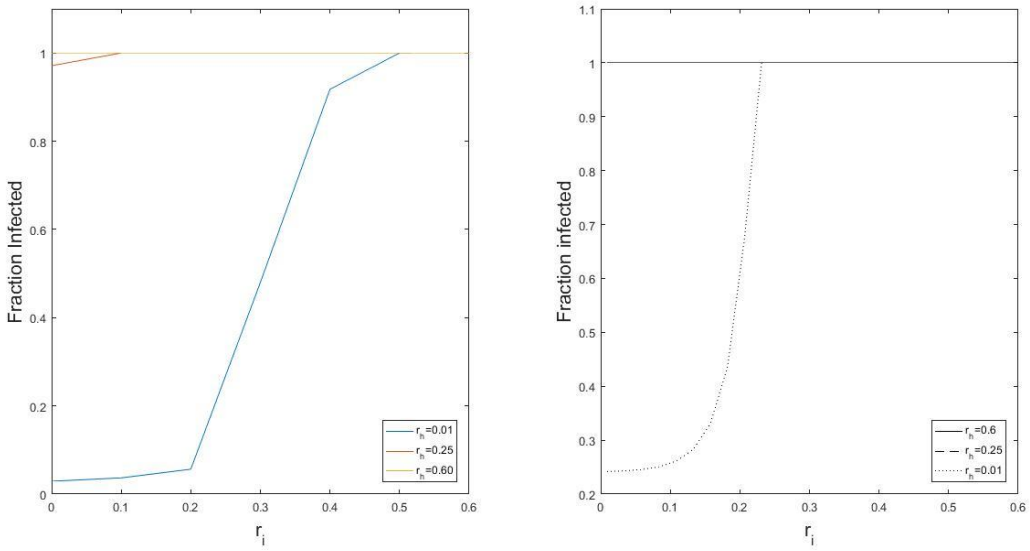

Figure 9: Left hand plot: Effect of the intrinsic growth rate of vectors on infected hosts (r_i, x-axis) on the fraction of hosts infected after 30 days (I, y-axis) at different growth rates of vectors on healthy hosts (r_h) for the IBM. Right hand side: Effect of the intrinsic growth rate of vectors on infected hosts (r_i, x-axis) on the fraction of hosts infected after 30 days (I, y-axis) at different growth rates of vectors on healthy hosts (r_h) using the ODE model code. Initial Conditions: 5% infected, parameter values found in the parameter table in Table 1.

We can see that an increase in r_i and r_h will increase the rate at which the virus spreads among the hosts in both models, decreasing the amount of time until the entire host population is infected (Figure 9). This relationship is what we would expect. If the intrinsic growth rate of vectors is increased, the probability of a vector reproducing increases. This then increases the number of vectors able to transmit the virus, increasing the probability and therefore the rate at which hosts become infected. This result agrees with the results of our sensitivity analysis, shown in Figure 5 for the IBM and Figure 8 for the ODE model. This analysis showed in both models that both r_h and r_i are inversely related to the time until 80% of the hosts become infected, meaning that an increase in r_h or r_i will decrease the time until 80% infection. Both r_h and r_i were considered significant based on their p-values for both models.

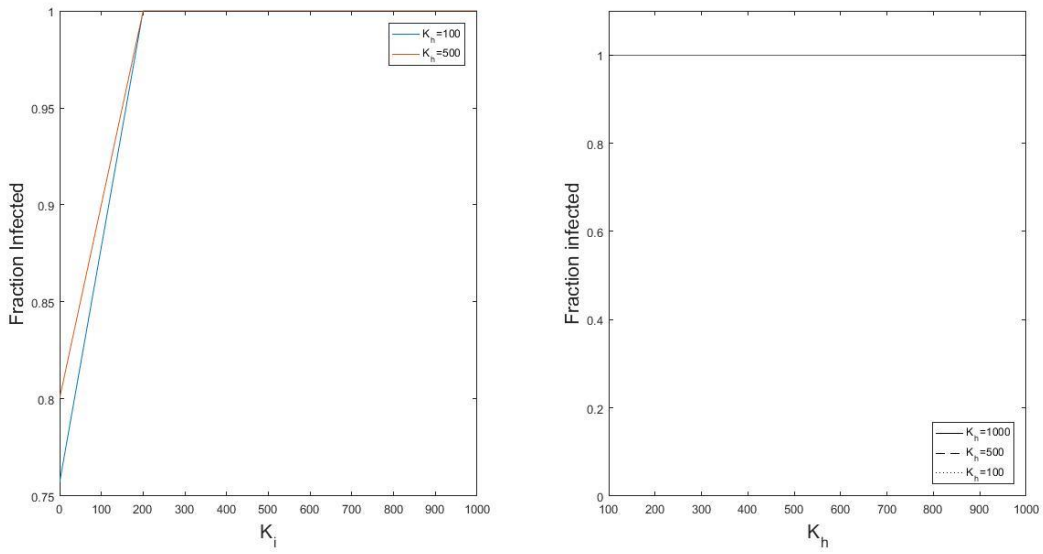

Figure 10: Left hand plot: Effect of the carrying capacity of healthy plants on the fraction of infected hosts after 30 days (x-axis) at different values for the carrying capacity of infected hosts (K_i, x-axis) for the IBM. Right hand plot: Effect of the carrying capacity of healthy plants on the fraction of infected hosts after 30 days (x-axis) at different values for the carrying capacity of infected hosts (K_i, x-axis) using the ODE model. Initial Conditions: 5% infected, parameter values found in the parameter table in Table 1.

We can see that the carrying capacity of healthy hosts and of infected hosts does not have a large effect on the fraction of hosts infected, as the fraction infected after 30 days is similar for all combinations of K_i and K_h that were tested (Figure 10). These results agree with our sensitivity analysis (Figure 5). In this analysis, we found that only K_h was considered statistically significant based on its p-value, and that its partial rank correlation coefficient showed a very weak relationship between it and the time until 80% of hosts were infected. In the ODE model, neither K_h nor K_i had significant p-values (Figure 8). This shows that K_h and K_i do not have a significant effect on the rate of infection of the host population in either model.

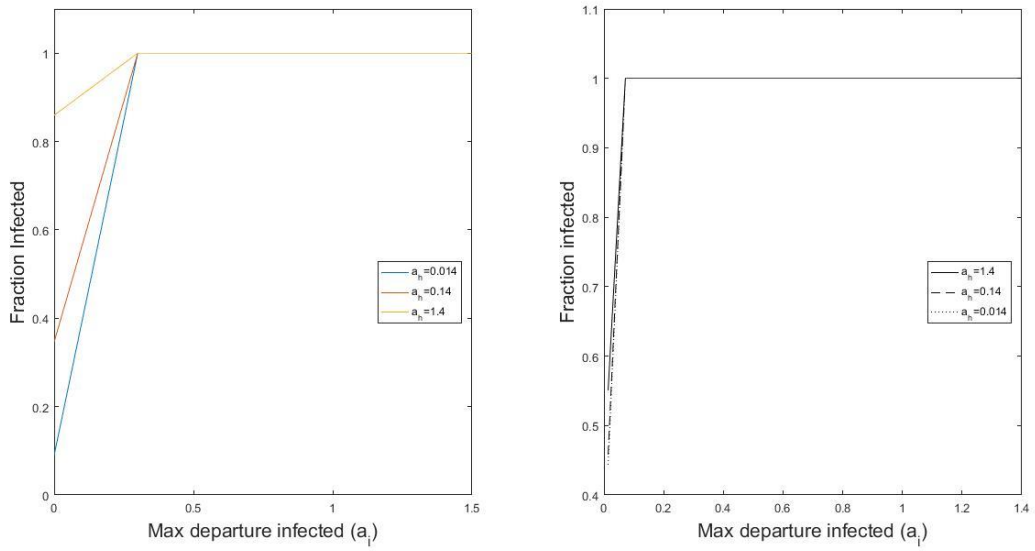

Figure 11: Left hand plot: Plot showing the fraction of hosts infected after 30 days as a function of the maximum departure rate from infected hosts (a_i, x-axis) at different a_h values for the IBM. Right hand side: Plot showing the fraction of hosts infected after 30 days as a function of the maximum departure rate from infected hosts (a_i, x-axis) at different a_h values using the ODE model. Initial Conditions: 5% infected, parameter values found in the parameter table in Table 1.

For the IBM, the effect of a_h and a_i on the fraction of hosts infected after 30 days is greater for values of a_i that are less than 0.5 (Figure 11). In this range, we can see that an increase in a_h leads to an increase in the fraction of hosts infected after 30 days, which corresponds to a lower time until all hosts are infected. For values of a_i greater than 0.5, all values of a_h yield 100 percent infection after 30 days. These results agree with the results of our sensitivity analysis in Figure 5. Both a_i and a_h are considered statistically significant, and show an inverse relationship to the rate of infection of the host population.

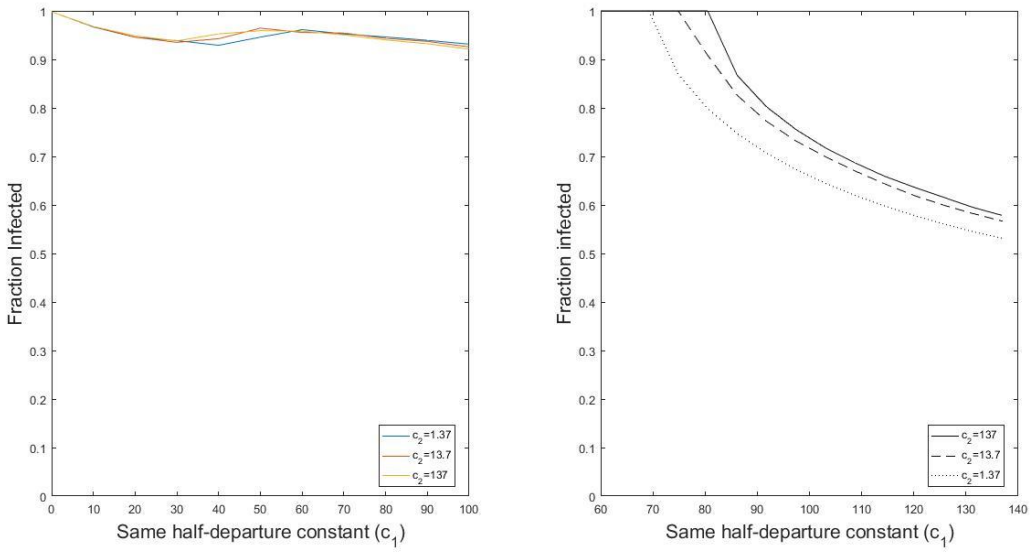

Figure 12: Left hand plot: Plot showing the fraction of hosts infected after 25 days as a function of the same half-departure constant (c_1) at different c_2 values for the IBM. Right hand side: Plot showing the fraction of hosts infected after 30 days as a function of the same half-departure constant (c_1) at different c_2 values using the ODE model. Initial Conditions: 5% infected, parameter values found in the parameter table in Table 1.

In Figure 12, we show the effect of half-departure constants on the fraction of infected hosts after 25 days in the IBM and in the ODE model. For the IBM, we can see that the value of c_2, the different status half-departure constant, has little effect on the fraction of hosts infected after 30 days. For the three values of c_2 tested, we see similar values for the fraction infected. However, we do see that an increase in c_1, the same status half-departure constant, leads to a decrease in the fraction infected after 30 days for all values of c_2 tested. In our sensitivity analysis, we found that neither c_1 nor c_2 was significantly related to the time until 80% of hosts become infected, meaning they do not have a large effect on the rate of infection of the host population.

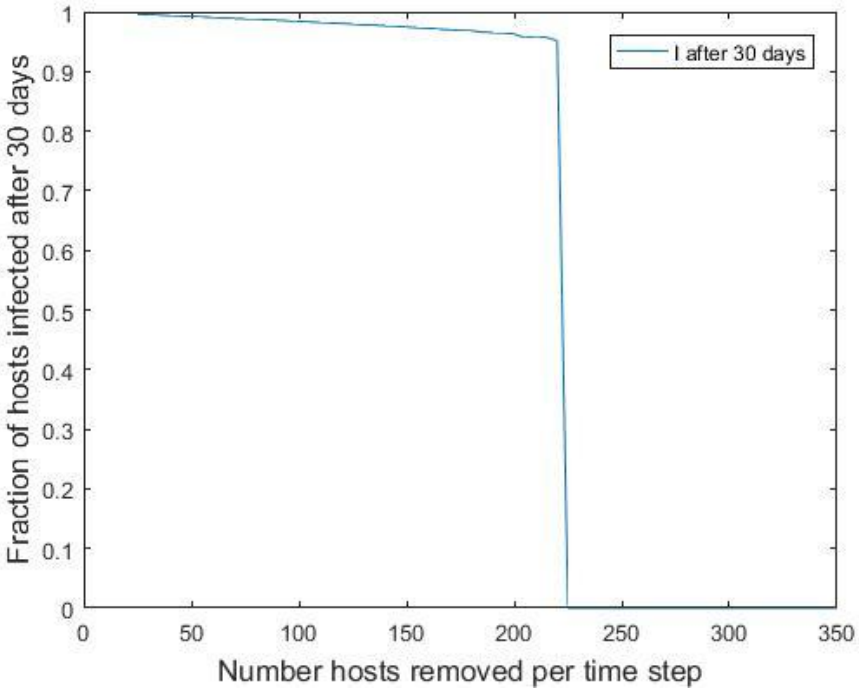

Figure 13: Graph showing the effect of removing a certain number of infected hosts at each time step on the fraction of infected hosts after 30 days.

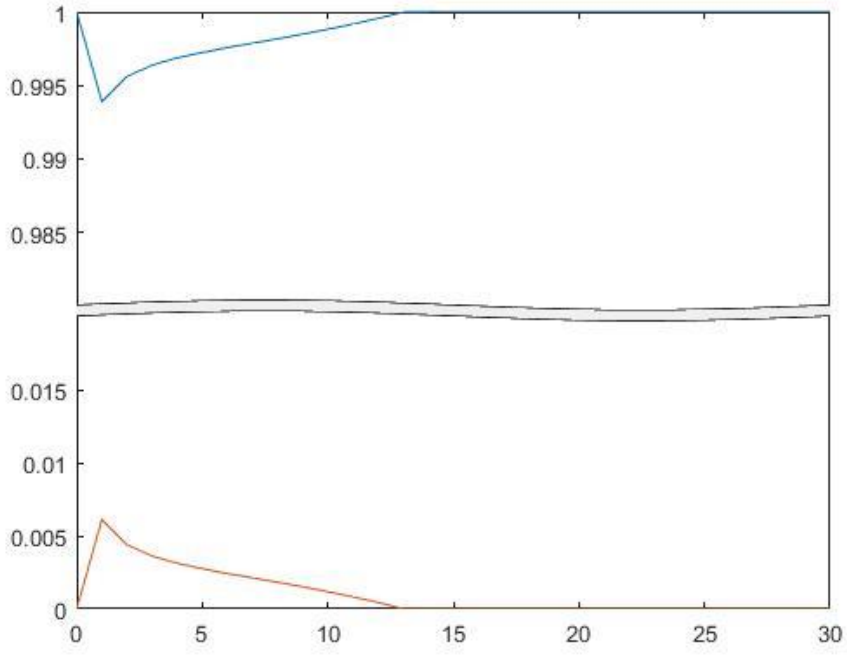

Figure 14: Graph showing the fraction of hosts infected at each time step if we remove 230 infected hosts at each time step.

To completely eradicate the pathogen from a population, approximately 230 infected hosts must be removed at each time step (Figure 13). We also notice that the fraction of infected hosts remains high (above 0.9) until we reach the removal of 225 hosts, at which point it decreases rapidly to 0 by the time we remove 230 hosts. We can also see that if we remove 230 infected hosts at each time step, the infection of host plants will end after 14 days (Figure 14).

Discussion and Comparison of Results to the ODE Model

The dynamics of the host population are slower in the IBM than in the ODE model (Figure 2). The right side of Figure 2 shows the fraction of healthy and infected hosts over time for the ODE model, ending when 100% of hosts become infected. The entire population becomes infected by 23 days. This is a faster spread than we see in the IBM, where total infection takes roughly 30 days. Due to the speed of this spread in the ODE model, we can see that the slope of the lines become nearly vertical as the infection as we approach 23 days. This is because the ODE model is a mean field model, which uses averages over the entire population of hosts and vectors to describe the spread of the virus among hosts and vectors. This means that, after a majority of the hosts become infected, infection of the entire population will occur quickly, as the rate of host infection is a constant rate that is dependent on the number of viruliferous vectors in the population, which in turn is determined by the number of infected hosts. Therefore, we see a positive feedback from the increase in the number of infected hosts over time, where an increase in the number of infected hosts increases the number of viruliferous vectors in the population, which then increases the fraction of infected hosts, and so on. This positive feedback occurs in the IBM, but at a much lower rate. This is because infection is probabilistic in the IBM. This stochasticity means that infection occurs at a slower rate than it would in a mean field model that does not have such stochasticity, and that the time until the final host is infected will

be longer than in a mean field model. The IBM also allows for the clumping of vectors on specific hosts, as shown in Figure 4, which would make the spread of the pathogen slower in the IBM than in the ODE model, which does not allow for clumping.

We see roughly the same vector population dynamics as we do in the ODE model (Figure 3). The major differences in the dynamics occur towards the end of simulations, as the fraction of hosts that are infected approaches 1. In the ODE model, we see that the N_i population is increasing in size at the end of the simulation, while in the IBM, we see that the size of the N_i population remains relatively constant at the end of the simulation. This behavior is possibly due to the stochastic elements introduced in the IBM, or due to the clumping of individuals on specific hosts. Though we see the expected drop to zero in the N_h and V_h populations as the fraction of healthy hosts reaches 0, the N_i population will take longer to respond to the complete infection of the host population because infection of vectors is probabilistic. This means that, though all non-viruliferous vectors are on infected hosts after 30 days, their infection at the proceeding time steps is not guaranteed. In the ODE model, we see that these interactions occur at a faster rate, accounting for the difference in population dynamics that we observe in the IBM.

In Figure 9, we see the effect of the growth rate of vectors on the fraction of infected hosts after 30 days for both the IBM and the ODE model. For the IBM, at the value r_h=0.01, we see that an increase in r_i leads to an increase in the fraction of hosts infected after 30 days. For the other values of r_h that were tested, we see that all values of r_i yielded total infection of the host population after 30 days. This pattern is similar in the ODE model. The qualitative relationship is the same, but we observe that, in the IBM model, lower values of r_i yield 100 percent infection of the host population than in the ODE model. The factors driving this difference are currently unclear. We do observe that, in the IBM, r_h has a much larger negative

PRCC value than r_i, but in the ODE model, we see the opposite relationship (Figure 5, Figure 8). This difference in the magnitudes of the PRCC values may be related to the slightly different dynamics that we observe as a result of clumping and other characteristics unique to the IBM, and therefore warrants further investigation.

We see very similar results when comparing the effect of the carrying capacity of infected and healthy hosts on the fraction of infected hosts after 30 days between the two models (Figure 10). In the IBM, K_i values less than 200 led to less than 100% host infection after 30 days, whereas in the ODE model, 100% infection is achieved after 30 days for all K_i and K_h values. Other than this small difference, we see the same pattern in our results. Between the two models, we also find that the carrying capacities are not considered statistically significant according to their p-values (Figure 5, Figure 8). These results tell us that in both the IBM and the ODE, the carrying capacities of K_h and K_i do not have a large effect on the fraction of infected hosts.

We can see that maximum departure constants seem to have more of an effect on the rate of host infection in the IBM than they do in the ODE model (Figure 11). For the ODE, we can see that for most values of a_i and a_h, host infection is 100% after 30 days, while this not true in the IBM model. We also observe that in the IBM, a_h has a inverse relationship with the time until 80% of hosts become infected, while in the ODE, a_h has an direct relationship with this output measure (Figure 5, Figure 8). The reason for these differences between models is currently unclear and also requires further investigation.

In Figure 12, we show the effect of half-departure constants on the fraction of infected hosts after 25 days in the IBM and in the ODE model. Between the two models, we see different behavior in terms of the shapes of the graphs. In the IBM, the fraction of hosts infected after 25

days never falls below 0.92 for any value of c_2, while in the ODE, for all value of c_2, the fraction of infected hosts falls to 0.53. However, despite this difference, we see that in both models, the value of c_2 has little visible effect on the fraction of hosts infected, while the higher values of c_1 appear to lead to a lower fraction of infected hosts after 30 days. This means that a lower value of c_1 appears to cause the time until total infection of the host population to increase. However, in the IBM, this relationship was not statistically significant (Figure 5), while in the ODE, this relationship was found to be statistically significant (Figure 8).

Conclusions

We have found that, despite quantitative differences in the results, the individual-based model and the mean field model yield qualitatively similar results in many areas. One major difference that we have observed between the two models is that the individual-based model caused slower population dynamics over time due to the introduction of stochastic components to the model. We have also found that individual-based models can have limitations due to lack of computing power or memory storage, which can limit the amount of investigation that can be done using an IBM. Additionally, using an individual-based model can introduce additional challenges in terms of sensitivity analysis because of the introduction of addition uncertainty to the system. However, despite these challenges, we see that the individual-based model produces roughly the same results as the mean field model in many areas. This means that the additional stochasticity of the individual-based model did not give us additional information on the system that was not already found using the mean field model. However, the individual-based model does allow us to explore the effects of some phenomena on the system that cannot be explored using a mean field model. This includes the effect of phenomena such as clumping on our results, as well as the investigation of individual-level changes to the system, such as the study of

possible control measures dealing with the removal of individuals from the system. This shows that both mean field models and individual-based models can be useful and informative modeling choices, offering insight into different dynamics in the same system.

Future Work

There are several areas of interest that require further investigation. In the parameter and PRCC simulations, we observed slightly different effects of parameters such as a_h, a_i, r_h, and r_i on the time until 80% of hosts became infected between the two models. In particular, we noticed a large difference in the significance of r_h and r_i to this output measure between the IBM and ODE model. The driving forces behind these differences, such as the effects of clumping of vectors on hosts, merit further study. Ways to study these effects include performing a PRCC analysis which does not include reproduction-related clumping of vectors on hosts and comparing it to the PRCC analysis which does include clumping, and to run simulations which vary departure and settlement rates of vectors to study how these changes affect the amount of clumping in the system. Another area for future work is further simulation of potential control measures for the spread of the pathogen, as well as the addition of further complexity, such as mortality traits, into the system.

Appendix

The vector dynamics and host dynamics equations for the Shaw et al. ODE model are shown below, for reference [6].

$$\frac{dN_h}{dt} = r_h[N_h + V_h]\left[1 - \frac{N_h + V_h}{K_h HF}\right] - \alpha_{nh} N_h I^{\delta} + \alpha_{ni} N_i\left[1 - I^{\delta}\right] - \beta_i \rho_{vh} N_h$$

$$\frac{dN_i}{dt} = \alpha_{nh} N_h I^{\delta} - \alpha_{ni} N_i\left[1 - I^{\delta}\right] + \beta_i \rho_{vh} N_h - \beta_v N_i + \gamma V_h$$

$$\frac{dV_i}{dt} = r_i[N_i + V_i]\left[1 - \frac{N_i + V_i}{K_i IF}\right] - \alpha_{vi} V_i H^{\varepsilon} + \alpha_{vh} V_h[1 - H^{\varepsilon}] + \beta_i \rho_{vh} V_h + \beta_v N_i$$

$$\frac{dV_h}{dt} = \alpha_{vi} V_i H^{\varepsilon} - \alpha_{vh} V_h[1 - H^{\varepsilon}] - \beta_i \rho_{vh} V_h - \gamma V_h$$

$$\frac{dH}{dt} = -\beta_i \rho_{vh} \qquad \frac{dI}{dt} = \beta_i \rho_{vh}$$

$$\rho_h = \frac{V_h + N_h}{HF + \xi} \quad \rho_h = \frac{V_i + N_i}{IF + \xi} \quad \rho_{vh} = \frac{V_h}{HF + \xi}$$

References

1. Durrett, R.; Levin, S. A. The Importance of Being Discrete (and Spatial). *Theor. Popul. Biol.* 1994, *46*, 363–394.
2. Detto, M.; Muller-Landau, H. C. Rates of formation and dissipation of clumping reveal lagged responses in tropical tree populations. *Ecology* **2016**, *97*, 1170–1181.
3. Railsback, S. F.; Grimm, V. Agent-Based and Individual-Based Modeling. In; 2012; pp. 3–13.
4. Bazghandi, A. Techniques , Advantages and Problems of Agent Based Modeling for Traffic Simulation. **2012**, *9*, 115–119.
5. Irwin, M. E. EPIDEMIOLOGY OF BARLEY YELLOW DWARF : A Study in Ecological Complexity. **1990**, 393–424.
6. Shaw, A. K.; Peace, A.; Power, A. G.; Bosque-Pérez, N. A. Vector population growth and condition-dependent movement drive the spread of plant pathogens. *Ecology*.
7. Ingwell, L. L.; Eigenbrode, S. D.; Bosque-Pérez, N. A. Plant viruses alter insect behavior to enhance their spread. **2012**.
8. Marino, S.; Hogue, I. B.; Ray, C. J.; Kirschner, D. E. *A Methodology for Performing Global Ucertainty and Sensitivity Analysis in Systems Biology*; 2008; Vol. 254.
9. Kiureghian, A. Der; Ditlevsen, O. Aleatory or epistemic? Does it matter? *Struct. Saf.* **2009**, *31*, 105–112.